УДК 873.01

Генаш Максим Геннадійович, Олійник Володимир Валентинович
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"
(Київ, Україна)

ЗАСТОСУВАННЯ НЕЙРОННИХ МЕРЕЖ АРХІТЕКТУРИ UNET, DEEPLABV3, PSPNET ДЛЯ СЕМАНТИЧНОЇ СЕГМЕНТАЦІЇ ОБЛИЧЧЯ НА ФОТОГРАФІЇ

Анотація. *У даній роботі досліджено можливість та доцільність застосування штучних нейронних мереж архітектури UNet, DeepLabV3, PSPNet для вирішення задачі семантичної сегментації обличчя на фотографії. Навчання мережі проводилося на датасеті Labeled Faces in the Wild (LFW) Part Labels Database. Семантична сегментація проводилася по 3 класам: волосся, область обличчя, фон. В результаті дослідження вдалося досягти достатньо високої точності сегментації для мережі UNet (Mean IoU = 85.6%, Pixel Accuracy = 95.7%), що відповідає рівню найкращих реалізацій моделей на датасеті LFW, при цьому досліджена модель достатньо компактна, завдяки чому може використовуватися у мобільних та веб-додатках.*

Ключові слова: *сегментація обличчя, штучні нейронні мережі, класифікація, LFW, UNet, Keras, обробка фотографій.*

Генаш Максим Геннадьевич, Олейник Владимир Валентинович
Национальный технический университет Украины
"Киевский политехнический институт имени Игоря Сикорского"
(Киев, Украина)

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ АРХИТЕКТУРЫ UNET, DEEPLABV3, PSPNET ДЛЯ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ЛИЦА НА ФОТОГРАФИИ

Аннотация. *В данной работе исследована возможность и целесообразность применения искусственных нейронных сетей архитектуры UNet, DeepLabV3, PSPNet для решения задачи семантической сегментации лица на фотографии. Обучение сети проводилось на датасете Labeled Faces in the Wild (LFW) Part Labels Database. Семантическая сегментация проводилась по 3 классам: волосы, область лица, фон. В результате исследования удалось достичь достаточно высокой точности сегментации для сети UNet (Mean IoU = 85.6%, Pixel Accuracy = 95.7%), что соответствует уровню лучших реализаций моделей на датасете LFW, при этом исследованая модель достаточно компактна, благодаря чему может использоваться в мобильных и веб-приложениях.*

Ключевые слова: *сегментация лица, искусственные нейронные сети, классификация, LFW, UNet, Keras, обработка фотографий.*

Maksym Henash; Ph.D., assistant professor Volodymyr Oliinyk
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
(Kyiv, Ukraine)

APPLYING UNET, DEEPLABV3, PSPNET NEURAL NETWORKS FOR SEMANTIC SEGMENTATION OF FACES ON PHOTO

Abstract. *This paper describes research on ability and feasibility of applying neural networks of UNet, DeepLabV3, PSPNet architectures in semantic segmentation of faces. The training was performed on Labeled Faces in the Wild (LFW) Part Labels Database. Semantic segmentation was performed by 3 classes: hair, face region, background. As the result of the research it was achieved fairly high level of segmentation accuracy for model UNet (Mean IoU = 85.6%, Pixel Accuracy = 95.7%) which is comparable with results of state of the art models on LFW dataset, meanwhile the trained model is compact enough to be appropriate for using in mobile and web applications.*

Keywords: *semantic face segmentation, neural networks, classification, LFW, UNet, Keras, photo processing.*

1. Вступ

Семантична сегментація зображень широко використовується у задачах комп'ютерного зору таких як пошук та виділення об'єктів, опис зображень, комплексне розуміння сцени та ін. Один із напрямків семантичної сегментації – сегментація обличчя, де задача полягає у правильному попіксельному назначенні класів елементам обличчя, таких як власне контур обличчя, ніс, рот, очі, волосся, тощо. Першими напрямками сегментації, які привернули увагу дослідників є семантична сегментація сцени й біомедичних зображень які можуть використовуватися, наприклад, для задач безпілотного керування транспортними засобами, чи автоматичної обробки біомедичних даних, таких як рентгенівські знімки. По зазначеним задачам навіть проводяться регулярні змагання, такі як PASCAL VOC challenge [1], що мотивує багатьох дослідників розвивати цю сферу.

Семантична сегментація обличчя довгий час привертала менше уваги, але останні роки завдяки розвитку індустрії доповненої реальності й підвищенні потужності процесорів мобільних пристроїв стала комерційно вигідною задачею для розважальних додатків із віртуального нанесення макіяжу [2] й перенесення стилю [3], редагування рис обличчя, обмін обличчями [4], зміни зачісок, кольору волосся, додавання різноманітних масок та ефектів, тощо. Крім того семантична сегментація обличчя може використовуватися як один із способів розпізнавання обличь [5, 6], визначення виразу/емоцій [7], передбачення явної особистості [8].

Семантична сегментація обличь має рад складностей у зв'язку із наявністю багатьох змінних умов, що варто брати до уваги: різний колір шкіри, освітленість, якість фотографії, поза та вираз обличчя, додаткові об'єкти на фоні, тощо. Особливу складність представляє сегментація волосся[9] через його різноманіття форм, кольорів, розмитість контурів та можливу збіжність по кольору з фоном.

Зважаючи на більший рівень дослідженості сфери сегментації сцени, для цієї задачі було реалізовано більше моделей, тож у даній роботі було перевірено можливість та доцільність застосування декількох кращих моделей реалізованих для сегментації сцени (PSPNet [10], DeepLabV3 [11]) та біомедичних знімків (UNet [12]) у задачі сегментації обличчя.

2. Пов'язані роботи

Останні роки штучні нейронні мережі та їх модифікації стали використовуватися для більшості задач штучного інтелекту. Семантична сегментація – не виключення. На зміну «класичним методам» [13] прийшли алгоритми засновані спершу на повнозв'яних нейронних мережах, згодом – засновані на згорткових нейронних мережах(таких як AlexNet [14], VGGNet [15], GoogLeNet [16]), які на даний момент дають найкращі результати у задачах пов'язаних із обробкою зображень.

Більшість робіт пов'язаних із семантичною сегментацією, що засновані на нейронних мережах використовують архітектуру автокодувальника із кодувальником – згортковою нейронною мережею й декодувальником – набором шарів оберненої згортки. У роботах останніх років було запропоновано підходи «розширеної згортки» [17] та «просторового пірамідального об'єднання» [11] для врахування контексту сцени.

У багатьох роботах також використовується Conditional random fields для пост-обробки результатів, наприклад [18, 19].

У роботах [18, 19, 20] наведено перевірку точності сегментації на датасеті LFW, тож у даній роботі результати порівнюються із цими роботами.

3. Особливості реалізації

3.1. Навчальні дані

Навчання проводилося на датасеті Labeled Faces in the Wild (LFW) Part Labels Database [21], адже це найбільший із публічно доступних датасетів який включає в собі 2927 пар фотографій лиця людей і відповідних їм результатів попіксельної сегментації. Фотографії зроблені для різних людей у різних позах, із різними об'єктами на фоні. На фотографіях сегментовані фон, шкіра обличчя (включаючи вуха і шию) й волосся(включаючи вуса й бороди за наявності).

Під час навчання усі зображення було масштабовано до розміру 64x64px для підвищення швидкості навчання, адже у межах дослідження розмір зображень був не суттєвим (для використання у реальних додатках можна провести повторне навчання нейронної мережі на більших зображеннях із налаштуваннями моделі які показали найкращі результати).

При подачі зображень на вхід моделі із генератора з кожним зображенням ще додатково виконувалися випадкові модифікації (з метою уникнення перенавчання): поворот до 20°, можливо відображення по горизонталі, масштабування до 20%.

Співвідношення тренувальної / валідаційної вибірки: 80% / 20% (зображення для відповідних вибірок обиралися випадковим чином).

3.2. Метрики

Під час навчання моделей у якості функції втрат використовувалася Categorical cross entropy. Вимірювання проводилось по метрикам: попіксельна точність (Precision) [22], Mean IoU (Jaccard Index) [23].

3.3. Моделі

Досліджувані моделі нейронних мереж було реалізовано мовою Python 3.6 на фреймворку Keras [24] із Tensorflow backend [25] у відповідності до опису зазначеного у [10, 11, 12].

Експериментальним шляхом було виявлено що найбільшу точність дає розмір порції навчання (batch size) = 16 для UNet й 128 для Deeplabv3 та PSPNet (також було перевірено batch size = 8, 32, 64, 128, 256, 512).

Експериментальним шляхом було виявлено що найбільшу точність дає оптимізатор Adadelta [26] (також було перевірено Nesterov accelerated gradient descent [27], Adam [28]).

Для навчання кожної із досліджуваних моделей використовувалося 500 епох.

4. Отримані результати

У результаті дослідження було виявлено що на датасеті LFW найкращі результати серед моделей PSPNet, DeepLabV3, UNet дає UNet. Порівняння отриманих результатів сегментації наведено у таблиці 1.

Розмір навченої моделі UNet становить 65.4 Мб.

Порівняння отриманих результатів із існуючими роботами наведено у таблиці 2.

Приклад сегментації обличчя на навченій моделі UNet наведено на рис. 1.

Рис. 1. Приклад результатів сегментації.

Таблиця 1. Порівняння результатів сегментації PSPNet, DeepLabV3, *UNet*

	Mean IoU, %	Precision, %
PSPNet	68.25	88.17
DeepLabV3	80.00	93.87
UNet	85.66	*95.73*

Таблиця 2. Порівняння отриманих результатів із існуючими роботами

	Mean IoU, %	Precision, %
[18]	н/д	92.47
[19]	88.82	96.67
[20]	н/д	94.82
UNet у цій роботі	85.66	*95.73*

5. Висновок

В результаті дослідження було виявлено, що серед моделей PSPNet, DeepLabV3, UNet найкращі результати для семантичної сегментації обличчя на фотографії при навчанні на датасеті LFW дає UNet. Причиною може бути те, що по-перше моделі PSPNet й DeepLabV3 мають значно більше шарів, тож для їх навчання необхідно більший датасет; по-друге моделі PSPNet, DeepLabV3 організовані таким чином щоб враховувати контекст сцени, який насправді є не дуже суттєвим у випадку сегментації фотографії де більшу частину займає обличчя людини.

Навчена модель UNet показала досить вдалі результати, що відповідає рівню найкращих реалізацій моделей на датасеті LFW, що означає можливість та доцільність її використання у задачах семантичної сегментації обличчя на фотографії, проте все ж існують роботи [29] на приватних датасетах які показують ще вищі результати (до 94.8% Mean IoU). Тож за необхідності підвищення точності сегментації необхідно вдосконалити датасет (більше тренувальних даних, точніше розмічення пікселів).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ:

1. Everingham M, Van GL, Williams CK, Winn J, Zisserman A. The Pascal visual object classes challenge. Int J Comput Vision 2010; 2: 303-338.
2. S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao, “Makeup like a superstar: Deep localized makeup transfer network,” CoRR, vol. abs/1604.07102, 2016.
3. M. Elad and P. Milanfar, “Style-transfer via texture-synthesis,” CoRR, vol. abs/1609.03057, 2016
4. I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast face-swap using convolutional neural networks,” CoRR, vol. abs/1611.09577, 2016.
5. F. Pujol, M. Pujol, A. Jimeno-Morenilla, and M. Pujol, “Face detection based on skin color segmentation using fuzzy entropy,” Entropy, vol. 19, p. 26, jan 2017.
6. K. Luu, C. Zhu, C. Bhagavatula, T. H. N. Le, and M. Savvides, “A deep learning approach to joint face detection and segmentation,” in Advances in Face Detection and Facial Image Analysis, pp. 1–12, Springer Nature, 2016.
7. S. L. Happy and A. Routray, “Automatic facial expression recognition using features of salient facial patches,” IEEE Transactions on Affective Computing, vol. 6, pp. 1–12, jan 2015.
8. Y. Güçlütürk, U. Güçlü, M. A. J. van Gerven, and R. van Lier, “Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition,” CoRR, vol. abs/1609.05119, 2016.
9. N. Wang, H. Ai, and S. Lao, “A compositional exemplar-based model for hair segmentation,” in Computer Vision – ACCV 2010, pp. 171–184, Springer Nature, 2011.
10. Zhao, Hengshuang, et al. "Pyramid scene parsing network." IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017.
11. Chen, Liang-Chieh, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611, 2018.
12. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

13. H. Zhu, F. Meng, J. Cai, and S. Lu, “Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation,” Journal of Visual Communication and Image Representation, vol. 34, pp. 12–27, 2016.
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25 (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.
15. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014.
16. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” CoRR, vol. abs/1409.4842, 2014.
17. F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” CoRR, vol. abs/1511.07122, 2015.
18. KHAN, Khalil, et al. Multiclass semantic segmentation of faces using CRFs. Turkish Journal of Electrical Engineering & Computer Sciences, 2017, 25.4: 3164-3174.
19. Güçlü, Umut, et al. End-to-end semantic face segmentation with conditional random fields as convolutional, recurrent and adversarial networks. arXiv preprint arXiv:1703.03305, 2017.
20. S. Saxena and J. Verbeek, “Convolutional neural fabrics,” CoRR, vol. abs/1606.02492v4, 2017
21. A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, “Augmenting CRFs with Boltzmann machine shape priors for image labeling,” in CVPR, 2013.
22. Csurka, Gabriela, et al. "What is a good evaluation measure for semantic segmentation?." BMVC. Vol. 27. 2013.
23. P. Jaccard. “Etude comparative de la distribution florale dans une portion des Alpe´s et du Jura. Bulletin de la Soci´et´e Vaudoise des Sciences Naturelles”, 37(142):547–579, 1901.
24. F. Chollet et al., “Keras,” https://github.com/fchollet/keras, 2015.
25. Abadi, M., Agarwal, A., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467, 2016.
26. Zeiler, Matthew D. "ADADELTA: an adaptive learning rate method." arXiv preprint arXiv:1212.5701, 2012.
27. Yu. Nesterov. A method of solving a convex programming problem with convergence rate o(1/k2). Soviet Mathematics Doklady, 27(2):372–376, 1983.
28. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.
29. Bazarevsky, V. and Tkachenka, A. “Mobile Real-time Video Segmentation”. [online] Google AI Blog. Available at: https://ai.googleblog.com/2018/03/mobile-real-time-video-segmentation.html, 2018